Research on Causes and Prediction of Global Warming Based on Mathematical Models

Xiaolin Jin[1,a], Changquan Huang[2,b,*]

[1]School of Accounting, Shanghai Lixin University of Accounting and Finance, Shanghai, 201620, China

[2]College of Artificial Intelligence, Yango University, Fujian, 350015, China

[a]landry105@163.com, [b]957762397@qq.com

*Corresponding author

Keywords: Global warming; One-sample t-test; ARIMA; LSTM; Information gain

Abstract: Global warming will cause little harm to the environment, which in turn will affect people's life and production. In this paper, we use time series models to predict the future global temperature change and explore the main causes of global warming through information gain. To address question 1, firstly, the monthly temperature from 2012-2021 was selected for a one-sample t-test with the temperature in March 2022, and it was obtained that the global temperature in March 2022 increased more than during the past decade. The ARIMA model and the LSTM model are used to predict the future global temperature level. For question 2, using the gray correlation analysis model, we found that there is a strong connection between global temperature and regions, and the southern hemisphere region has a greater influence on global temperature. Then, by establishing a multiple regression model to determine the relationship between global temperature and natural disaster factors. Finally, by calculating the indicators of carbon dioxide, forest area, and population size through information gain, it is concluded that the main cause of global warming is due to the excessive emission of carbon dioxide. In response to question three, in the future, we should reduce the use of energy sources such as coal and use more green and clean energy sources, thus reducing carbon dioxide emissions. Plant trees and increase forest cover, thus effectively curbing global warming.

1. Problem Restatement

(1) Explore whether the global temperature will increase more in March 2022 than it has during the past 10-year period, and then develop two or more mathematical models to describe past global temperatures and predict whether global temperatures will reach 20 degrees Celsius in 2050 and 2100, and determine which prediction model is most accurate.

(2) Explore whether there is a relationship between global temperature and time and location, analyze the effects of natural disasters on global temperature, and explore the main causes of global warming, then give ways to curb or mitigate global warming.

(3) Write a non-technical article to the APMCM organizing committee about the team's findings and suggestions for the future.

2. Problem Analysis

2.1 Analysis of Problem 1

(1) The first question, because the temperature in the tropics has not changed much within the last 10 years and is not sensitive to global warming, while temperate regions are more sensitive to global warming, the monthly average temperature in the north and south temperate zones is selected to measure whether the temperature in March 2022 has increased more than the last decade, using a one-sample t-test to verify whether there is a difference between the two.

(2) In the second question, firstly, the average temperatures of the southern and northern temperate zones from January 1899 to 2022 were selected as a measure of global temperature levels, and the

  DOI: 10.25236/icmmct.2023.010

time series data were pre-processed and found to be free of outliers and temporal misalignments as imagined. levels, and finally, adjust the parameters of the LSTM model using a genetic algorithm to improve the accuracy of the model.

(3) In the third question, the ARIMA model and the LSTM model are projected for 2050 and 2100, and if neither reaches 20 degrees Celsius, the length of the projection is continued to increase until it reaches 20 degrees.

(4) In the fourth question, the goodness-of-fit indicates the degree of fit between the prediction curve and the true curve, and if it is close to 1, the model predicts accurately.

2.2 Analysis of Problem 2

(1) In the first question, the relationship between time and global temperature is firstly studied by making a series graph of time and global temperature, and then the relationship between regional and global temperature is studied by observing the series graph of temperature, and then the gray correlation analysis is used to study the relationship between regional and global temperature.

(2) In the second question, the relationship between global temperature and natural disaster factors is investigated by building a multiple regression model.

(3)The third question, is by building an information gain model to determine the degree of influence of indicators such as carbon dioxide forest area population size on global temperature, so as to find out the main causes of global temperature warming.

(4) In the fourth question, the factors that mainly affect global temperature are analyzed in the third question in order to curb or mitigate global warming.

2.3 Analysis of Problem 3

Write a paper to the organizing committee about the team's findings as well as make some suggestions.

3. Problem Assumptions

1) There will be no sudden sharp rise and fall in global temperature

2) The average temperature of the north and south temperate zones replaces the global temperature

4. Symbol Description

Table 1 Symbol Description

Symbols	Meaning
t	T-value
$\bar{x}$	Average value
μ_0	The temperature in March 2022
H_0, H_1	Original hypothesis, the alternative hypothesis
y	Global Temperature
R^2	Goodness of fit
r	Gray correlation
g	Information Gain

The meaning of the symbols in the following paper is described in Table 1.

5. Model Building and Solving

5.1 Modeling and Solving Problem 1

Since tropical regions have little temperature variation throughout the year and are not sensitive to the effects of global warming, while temperature regions are more sensitive to the effects of temperature on global warming, this paper uses the average temperature of the north and south

temperate regions to measure global temperature and thus explore whether global temperatures are warming.

5.1.1 One-sample T-test

The first sub-question in question 1 is to investigate whether the global temperature increase in March 2022 is greater than that during the past decade. Therefore, the monthly temperatures in the north and south temperate zones from 2012-2021 are selected as a control against the average temperature in January-October 2022, and a one-sample t-test is used to investigate whether the change is greater [1].

The one-sample t-test is used to analyze whether there is a significant difference between the quantitative data and a determined value.

$$t = \frac{\bar{x} - \mu_0}{s_x} \quad (1)$$

H_0:Global temperatures in March 2022 will not rise more than they have in the past decade

H_1:Global temperatures to rise more in March 2022 than in the past decade

Where $\bar{x}$ denotes the mean of the sample, μ_0 denotes the average temperature of 17.068 degrees Celsius in January-October 2022, s_x denotes the variance of the sample, H_1 denotes the original hypothesis, and h1 denotes the alternative hypothesis.

Equation (1) was solved using SPSSPRO, and the following table was obtained.

Table 2 Table of results of one-sample t-test

Test value	Sample size	Average value	Standard deviation	t	P
17.068	120	16.051	2.395	-4.654	0.000***

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively

By looking at Table 2, the test value of 17.068 is the average temperature for January-October 2022, indicated as a sample size of 120 indicates the monthly temperature within the last decade. The significance p-value of 0.000***, which is less than 0.05, indicates that the original hypothesis is rejected and the alternative hypothesis is accepted, which is that the global temperature in March 2022 is greater than the increase within the last 10 years period.

5.1.2 ARIMA Model

The time series data should first be subjected to a preliminary check to determine whether there are any breaks in the time series and to prevent any time mismatch [2]. In this paper, by selecting the average temperature of the southern temperate zone and the northern temperature zone from January 1899 to October 2022, and there are no vacant values as well as time misalignment.

A time series is a sequence of values of an indicator arranged in chronological order. Time series analysis can be broadly divided into three main parts, which are describing the past, analyzing the law and predicting the future.

Time series forecasting model is to use the processing of the time series of the forecast target itself to analyze its trend. A time series usually has a superposition or coupling of the following forms of change.

(1) Changes in long-term trends.

(2) Seasonal variations.

(3) Cyclic variation.

(4) Irregular variation.

The variation term is used to represent the long-term trend, the seasonal variation trend term, the cyclic variation trend term, and the random error or random disturbance.

ARIMA(p,d,q) model

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} \quad (2)$$

The first prediction model developed in this paper is ARIMA(p,d,q), by which the prediction describes the past global temperature and predicts the future global temperature level.

The ARIMA model was solved using SPSSPRO and the following graphs were obtained.

Table 3 Time series smoothness test

ADF Inspection Form							
Variables	Difference order	t	P	AIC	Threshold		
					1%	5%	10%
Global Temperature	0	0.999	0.994	1302.635	-3.435	-2.864	-2.568
	1	-14.792	0.000***	1298.482	-3.435	-2.864	-2.568
	2	-16.567	0.000***	1436.352	-3.435	-2.864	-2.568
Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively							

The ARIMA model requires the data to be smooth, and the original data are first checked for smoothness, and if the data are not smooth, they are differenced. ADF test was performed on the original data to judge its stability, and as can be seen from Table 3, when the difference was divided into 0th order, the significance P-value was 0.454, which was greater than 0.05, indicating that the original hypothesis (unsteady time series) was accepted. Continuing its differencing, when the difference is divided into 1st order, the P value is 0.000**, which is less than 0.05, indicating the acceptance of the alternative hypothesis (smooth time series), so when the first order differencing, the change series becomes a smooth time series.

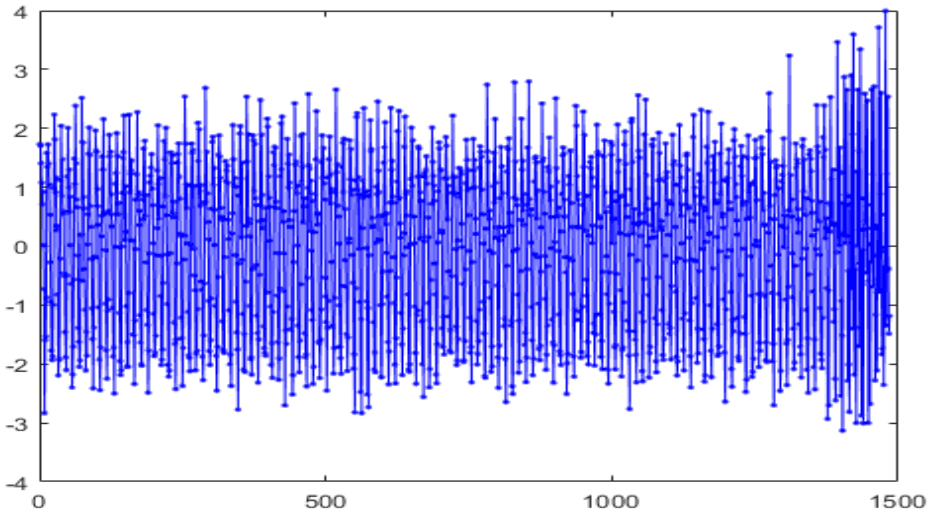

Figure 1 First-order differential time series

As can be seen from Figure 1, when the time series becomes smooth after the first-order differencing, the ARIMA model can be fitted with the first-order differencing data at this time.

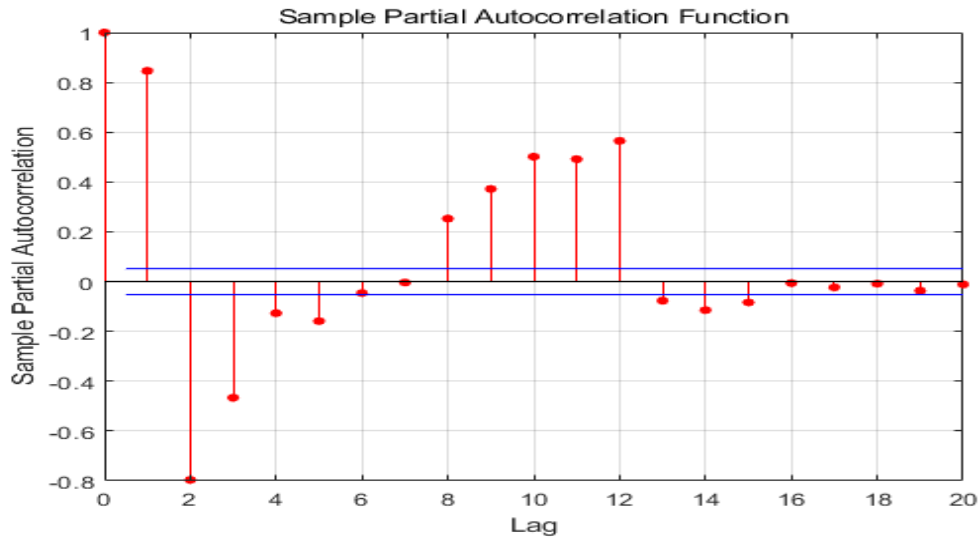

Figure 2 PACF

As can be seen from Figure 2, the truncation of the tail occurs when at orders 4-6, indicating that MA(p) = 4, 5, 6 at this time.

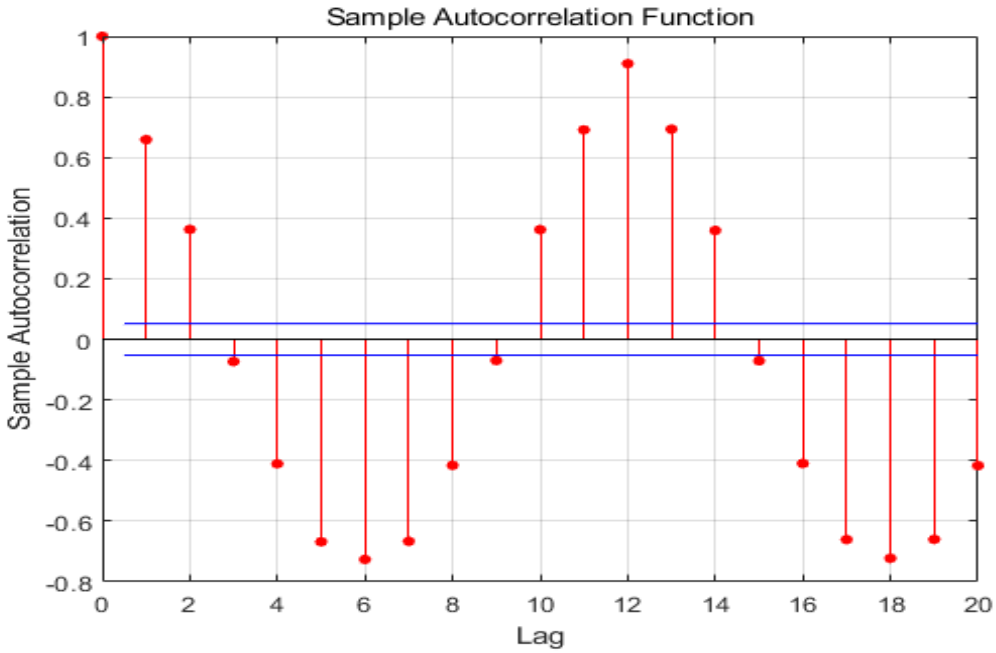

Figure 3 ACF

From Figure 3, it can be seen that the trailing phenomenon occurs when at order 5-7, indicating that AR(q) = 5, 6, 7.

The values of ARIMA(p,d,q) were determined by SPSSPRO and the following table was obtained.

Table 4 ARIMA model test table

ARIMA model (5,1,4) test table		
	Symbols	Value
Information Guidelines	AIC	2074.606
	BIC	2132.941
Goodness of fit	R^2	0.964
Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively		

According to the AIC and BIC criteria, the average minimum of AIC and BIC is obtained from Table 4 when p=5 and q=4. The goodness of fit at this time is 0.964, indicating a good fit, so the final

model is chosen as the ARIMA (5,1,4) model, as follows.

$$y_t = 0.001 + 1.187y_{t-1} - 0.005y_{t-2} - 0.402y_{t-5} - 1.688\varepsilon_{t-2} + 1.598\varepsilon_{t-3} - 0.905\varepsilon_{t-4} \quad (3)$$

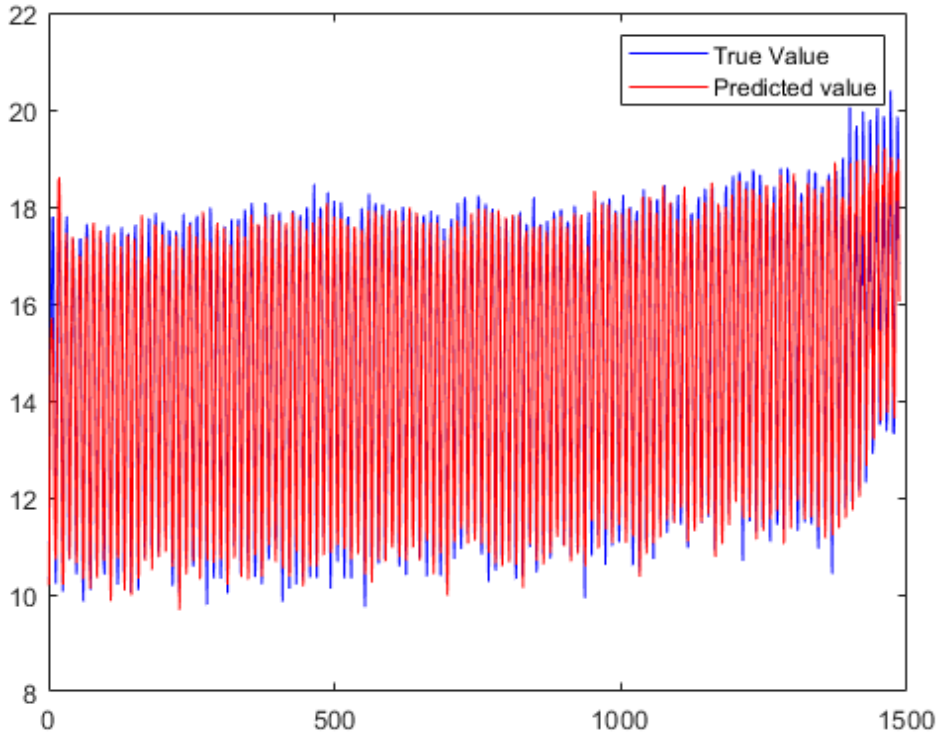

Figure 4 Fitting graph

As can be seen from Figure 4, the prediction curve of the ARIMA(5,1,4) model fits well with the true curve, which indicates that this model fits well and the best fit is achieved in the middle time period.

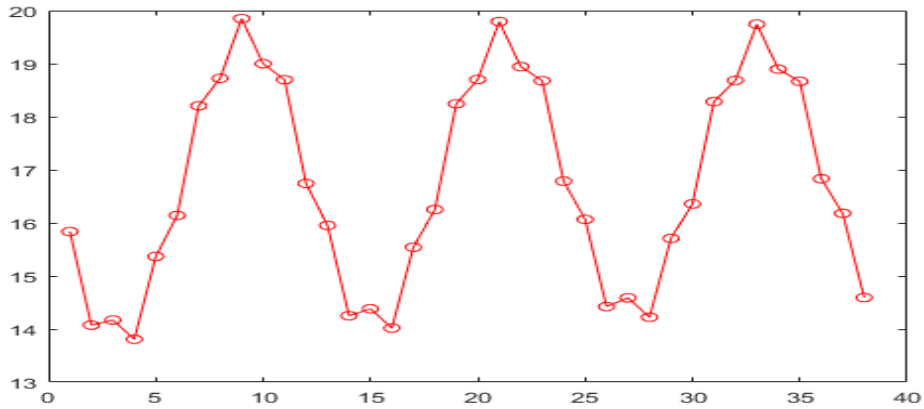

Figure 5 Global temperature level in the next 3 years

As can be seen in Figure 5, the temperature in November 2022 is 15.8, showing a decreasing trend. In the next 2023-2025, the monthly maximum temperature will exceed 20 every month, and the temperature of each year will show a cyclical change.

5.1.3 LSTM Model

LSTM is a special form of the recurrent neural network, which has a "gate" structure and will not be affected by weights, gradient disappearance, and explosion imagination; the LSTM model makes the network structure more perfect, faster convergence, and high accuracy of fitting and prediction [3].

The LSTM has three structures, which are the forgetting gate, the input gate, and the output gate, and the specific network structure is as Figure 6.

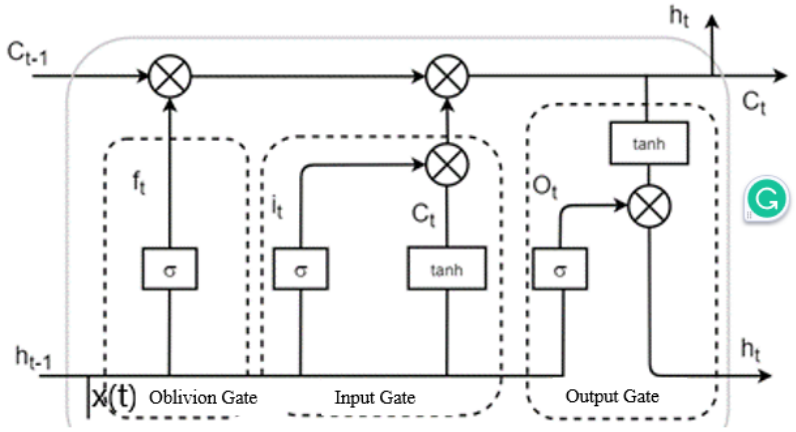

Figure 6 LSTM structure diagram

Oblivion Gate: Control whether every moment of information will be forgotten by the system

$$f_t = \sigma(W_t h_t + b_f) \tag{4}$$

Input Gate: Determine how much new information has been added to the current cell

$$i_t = \sigma(W_t[h_t, x_t] + b_i) \tag{5}$$

Unit: Activation Unit

$$C_t = \tanh(W_c[h_t, x_t] + b_c) \tag{6}$$

$$C_t = f_t C_{t-1} + i_t C_t$$

Output Gate: Determine whether messages are output by the system at each moment

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{7}$$

Final Output:

$$h_t = O_t \tanh(C_t) \tag{8}$$

The time series data were first normalized, and then 70% of the data were used as the training set and 30% as the test set, and the LSTM model was solved using Matlab to obtain the following figure.

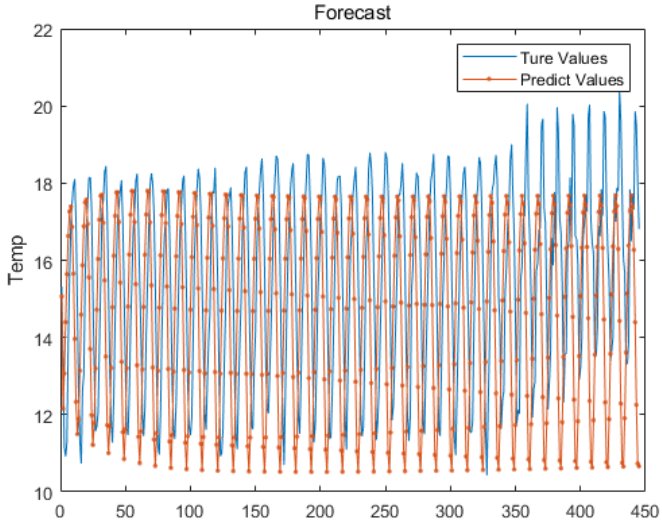

Figure 7 Test set true set and predicted values

From Figure 7, it can be seen that the prediction curve of the LSTM model on the test set does not fit well with the true curve, so the LSTM model is optimized using the particle swarm algorithm and the following graph is obtained.

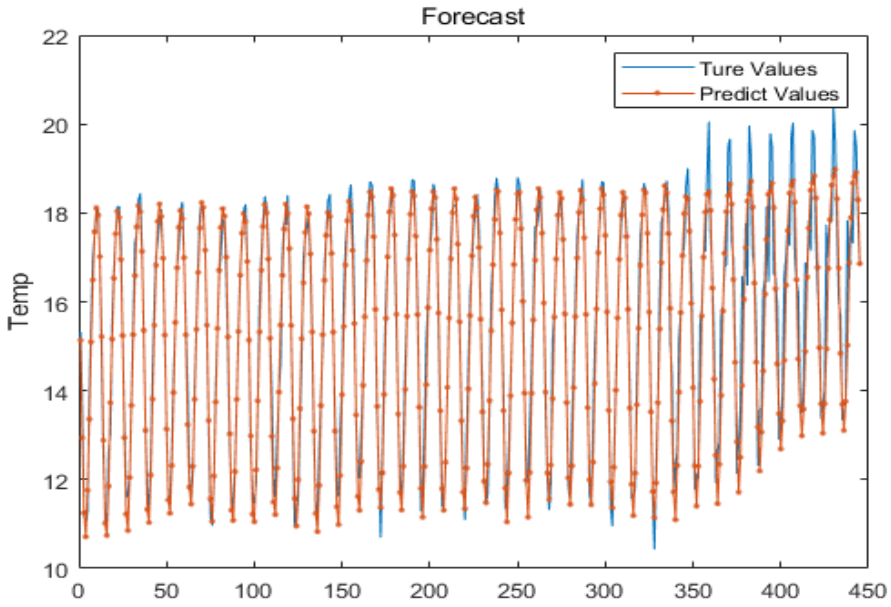

Figure 8 Optimization diagram of particle swarm algorithm

As can be seen from Figure 8, when the LSTM hyperparameters are adjusted using the particle swarm algorithm, the prediction curves fit the true curves very well, indicating that after optimization, the accuracy of the LSTM model for predicting global temperature levels is greatly improved.

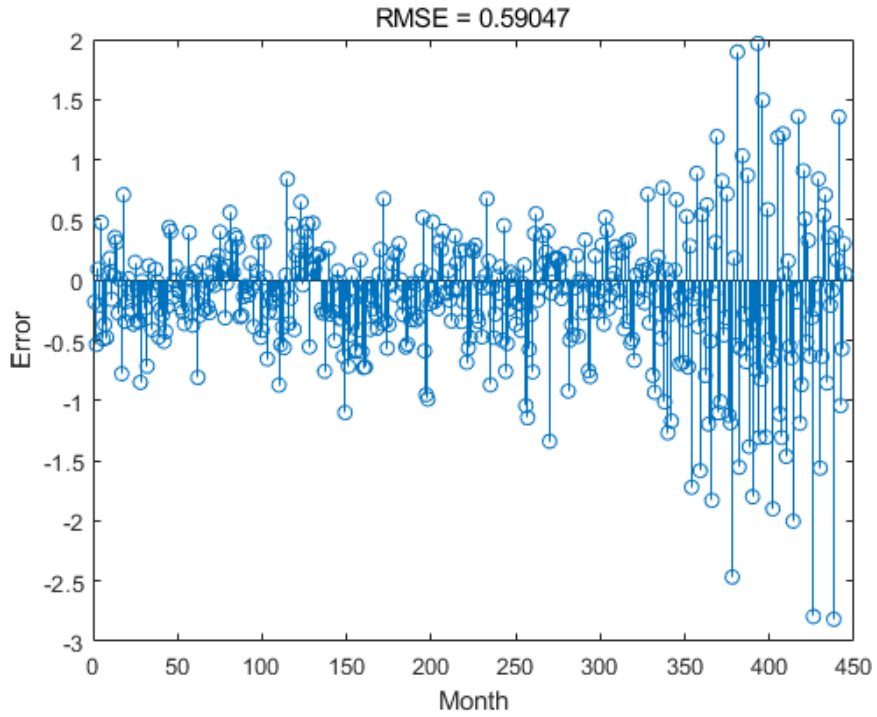

Figure 9 Error diagram

The closer the error is to 0, the better the model effect is. As can be seen from Figure 9, the error of the test set is very close to 0, which indicates that the accuracy of the LSTM model is greatly improved after particle swarm optimization.

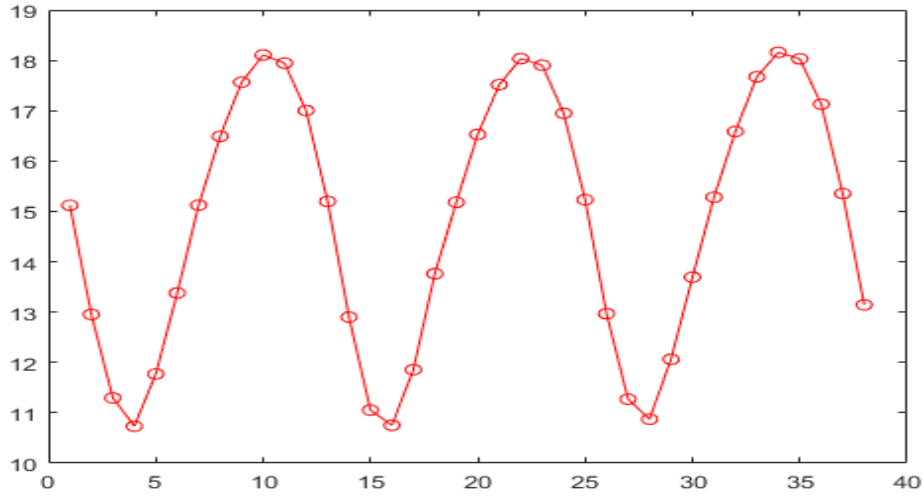

Figure 10 Global temperature levels over the next 3 years

As can be seen in Figure 10, the global temperature level in November 2022 is 15.1, flat and on a decreasing trend. In the next three years, the global temperature level shows a certain cyclical variation.

5.1.4 Predictions for 2050 and 2100

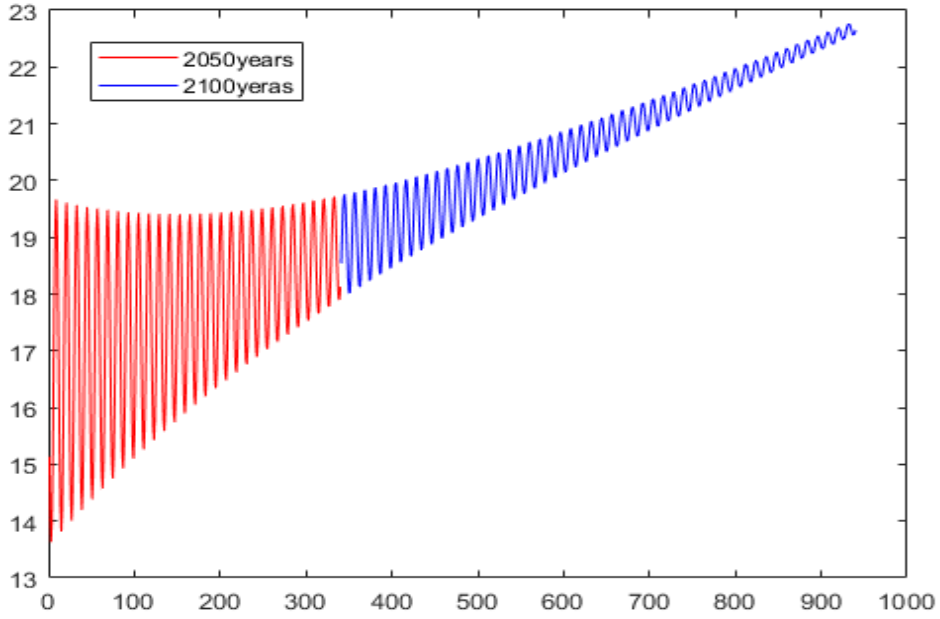

Figure 11 ARIMA-2050 and 2100

As seen in Figure 11, the ARIMA model predicts that the global temperature level does not reach 20 degrees Celsius in 2050. In contrast, it is already well above 20 degrees Celsius at 2100, reaching 22.65 degrees Celsius, and has already reached 20 degrees Celsius in 2059.

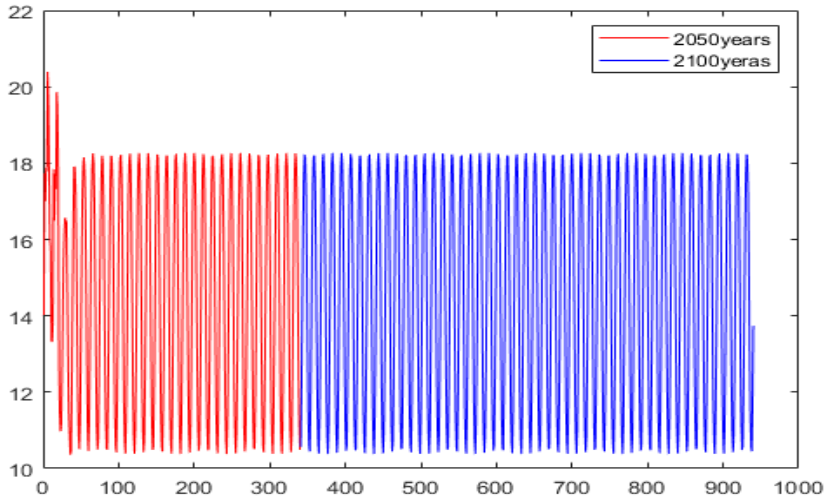

Figure 12 LSTM - 2050 and 2100

From Figure 12, it can be seen that in 2050 and 2100, instead of reaching 20 degrees Celsius, 20 degrees Celsius is reached already in 2023.

5.1.5 Optimal Model

The goodness-of-fit is a value between 0 and 1. If it is closer to 1, it means that the model fits better and the model is more accurate, so the goodness-of-fit is chosen as the judge of the accuracy of the model, which is calculated as follows.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9}$$

According to equation (16), the goodness of fit of ARIMA model and LSTM model are calculated respectively, and the goodness of fit of ARIMA is 0.964, while the goodness of fit of LSTM is only 0.9, which indicates that the prediction effect of ARIMA model is more accurate.

The residual is the difference between the true value and the predicted value. If the residual is closer to 0, it indicates that the predicted value is closer to the true value, so the residual can also be used to judge the accuracy of the model.

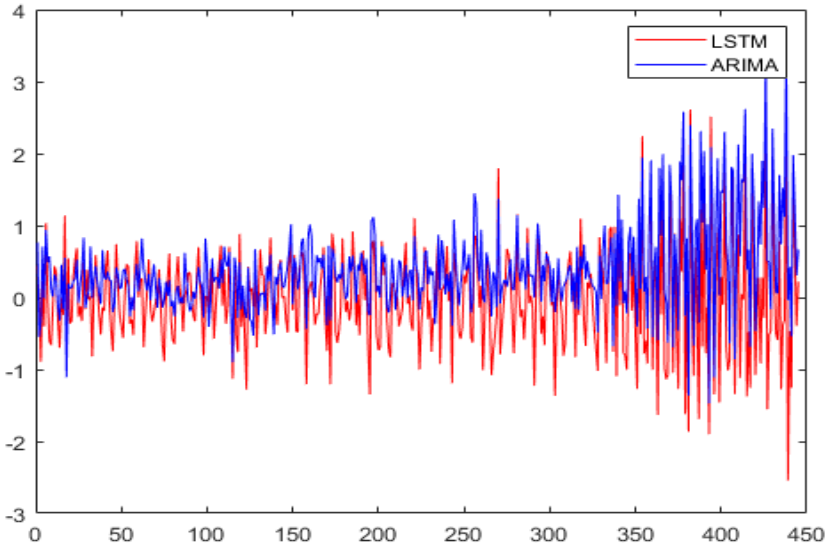

Figure 13 Residuals of ARIMA and LSTM

As can be seen from Figure 13, the residuals of the ARIMA model are more concentrated in the 0 annex, which indicates that ARIMA has better prediction.

5.2 Modeling and Solving Problem 2

5.2.1 Time and Global Temperature Trends

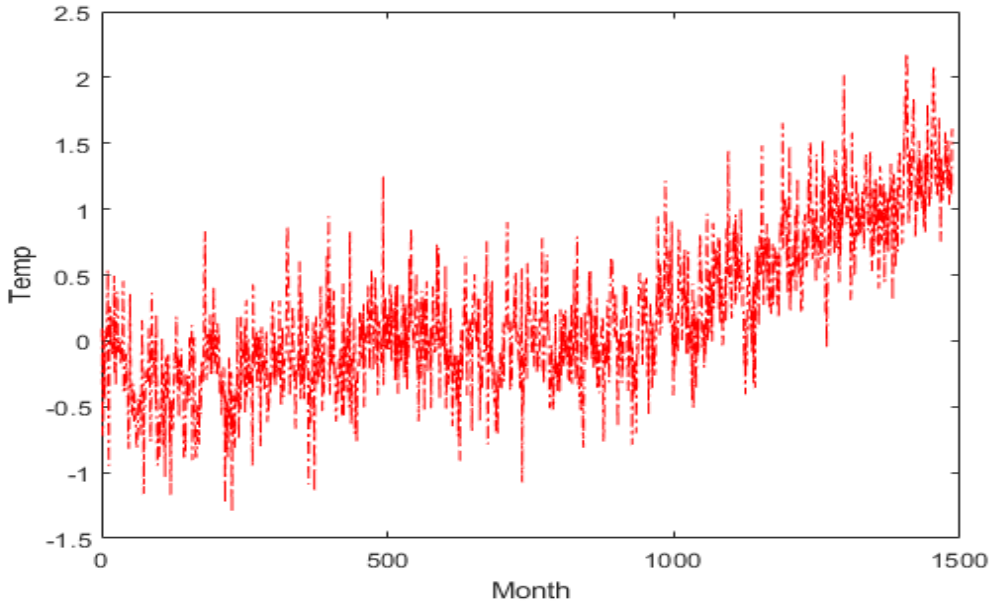

Figure 14 Time vs. global temperature trend

As can be seen from Figure 14, the global temperature is increasing over time, with a small increase in the early part of the period, but an increasing increase from 1982 onwards. From 2004 onwards, the global temperature does not fall below 0 degrees, indicating that the global temperature is increasing the later it gets.

5.2.2 Gray Correlation Analysis

Gray correlation analysis is based on the similarity of the geometry of the data series curves to determine whether the child series are close to the parent series, such as if the closer the curves are, the greater the degree of correlation between the corresponding series, and vice versa, the smaller [4]. The degree of correlation between southern hemisphere and northern hemisphere temperatures and global temperatures is established by gray correlation analysis to determine the relationship between global temperatures and regions.

Parent sequence: A data series reflecting the behavioral characteristics of the system, corresponding to the dependent variable Y

$$Y = [y_1, y_2, ..., y_m]^T \quad (10)$$

Subsequences: A data series consisting of factors affecting the behavior of the system, corresponding to the independent variable X

$$X_{nm} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \quad (11)$$

The number of gray correlation coefficients between the ith subsequence in the calculation and the parent sequence Y at the kth sample.

$$\xi_i(k)=\frac{\min\limits_{s}\min\limits_{t}|y_t-x_{st}|+\rho\max\limits_{s}\max\limits_{t}|y_t-x_{st}|}{|y_t-x_{it}|+\rho\max\limits_{s}\max\limits_{t}|y_t-x_{st}|} \tag{12}$$

Where is the discrimination coefficient, the gray correlation of the ith subsequence with the parent sequence is

$$r_i=\sum_{k=1}^{n}\omega_t\xi_i(k) \tag{13}$$

The grey correlation analysis model was solved using SPSSPRO and the following graphs were obtained.

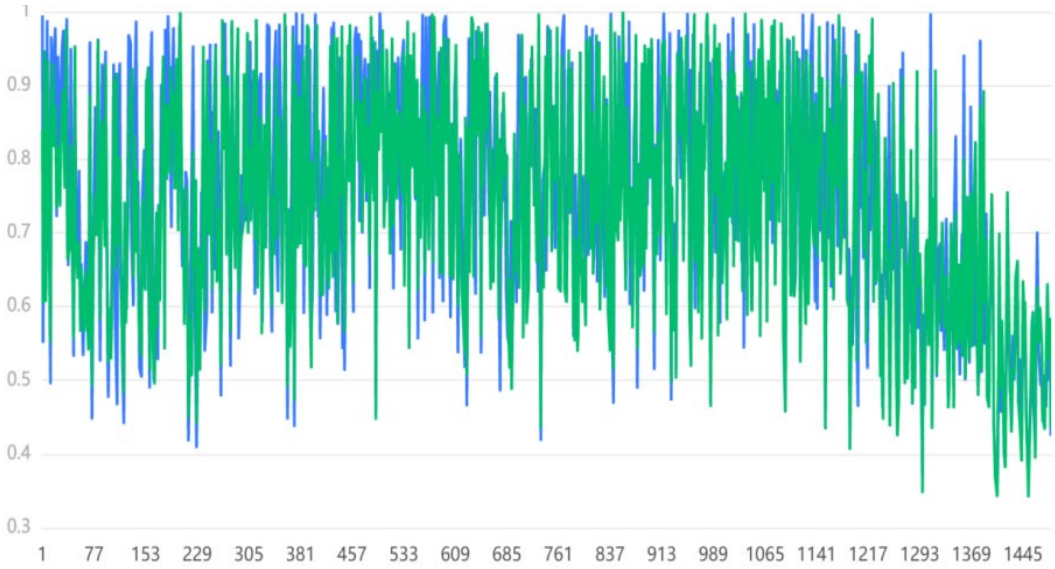

Figure 15 Correlation coefficient map (blue - southern hemisphere, green - northern hemisphere)

A larger value of correlation coefficient indicates that the subsequence of this sample is more strongly correlated with the parent sequence. As can be seen from Figure 15, the correlation coefficient of the Southern Hemisphere temperature is generally larger than that of the Northern Hemisphere temperature, indicating that the global temperature is more influenced by the temperature in the Southern Hemisphere.

Table 5 Gray correlation degree

Subsequences	Relevance	Ranking
Southern Hemisphere	0.746	1
Northern Hemisphere	0.734	2

As can be seen from Table 5, the correlation of the average temperature in the southern hemisphere is 0.746 higher than that in the northern hemisphere by 0.012, which indicates that the global temperature is influenced by different regions and is more influenced by the temperature in the southern hemisphere.

5.2.3 Multiple Regression Model

Australia has frequent and long-lasting large-scale forest fires, so the average monthly temperature zone of Australia in the calendar year was chosen to measure forest fires as a natural disaster factor. Similarly, the temperature in the United States was chosen as the COVID-19 natural disaster factor and the temperature in Indonesia as the volcanic eruption natural disaster factor. A multiple regression model is constructed between these three natural disaster factors and global temperature to analyze the impact of natural disasters on global temperature.

$$y=a_0+a_1x_1+a_2x_2+a_3x_3+\varepsilon \tag{14}$$

Where y denotes global temperature, x_1、x_2、x_3 denotes natural disaster factors such as volcanic eruption, forest fires, COVID-19, a_0 denotes constant term, and ε denotes random error.

The multiple regression model was solved using SPSSPRO and the following graphs were obtained.

Table 6 Multiple regression analysis

	Non-standardized coefficient		Standardization factor	t	P	VIF	R^2	F
	B	Standard error	Beta					
Constants	-3.176	0.525	-	-6.053	0.000***	-	0.503	F=495.146 P=0.000***
Forest fires	0.17	0.005	1.139	34.866	0.000***	3.156		
COVID-19	0.066	0.002	0.886	27.106	0.000***	3.161		
Volcanic eruptions	0.333	0.019	0.323	17.54	0.000***	1.004		

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively

Judgment of whether the multiple regression model is valid or not mainly depends on whether it passes the F-test, while in some cases, there is no need to pay much attention with the goodness-of-fit. As can be seen from Table 6, in the F-test, a significant P-value less than 0.05 indicates that the F-test is passed, indicating that the multiple regression model is valid and the final model is obtained as follows.

$$y = -3.176 + 0.33x_1 + 0.17x_2 + 0.066x_3 + \varepsilon \tag{15}$$

From equation (20), it can be seen that among the coefficients of the three natural disaster factors, the coefficient of volcanic eruption is the largest at 0.33, and the smallest is COVID-19, which indicates that volcanic eruption has the largest impact on global temperature, while COVID-19 has a smaller impact on global temperature.

5.2.4 Information Gain Model

The information gain indicates the degree to which the information of feature X reduces the uncertainty of the information of Y [5]. That is, the information gained can determine which feature has a greater influence on the independent variable Y. Thus, the degree of influence of features such as population size, carbon dioxide, forest area, volcanic eruption, and COVID-19 on global warming can be calculated.

Step1: Calculate the empirical entropy of data set D

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \tag{16}$$

Step2: Calculate the empirical entropy H(D|A) of the features on the data set D

$$H(D \mid A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \tag{17}$$

Step3: Calculate the information gain

$$g(D, A) = H(D) - H(D \mid A) \tag{18}$$

The solution was solved using SPSSPRO and the following figure was obtained.

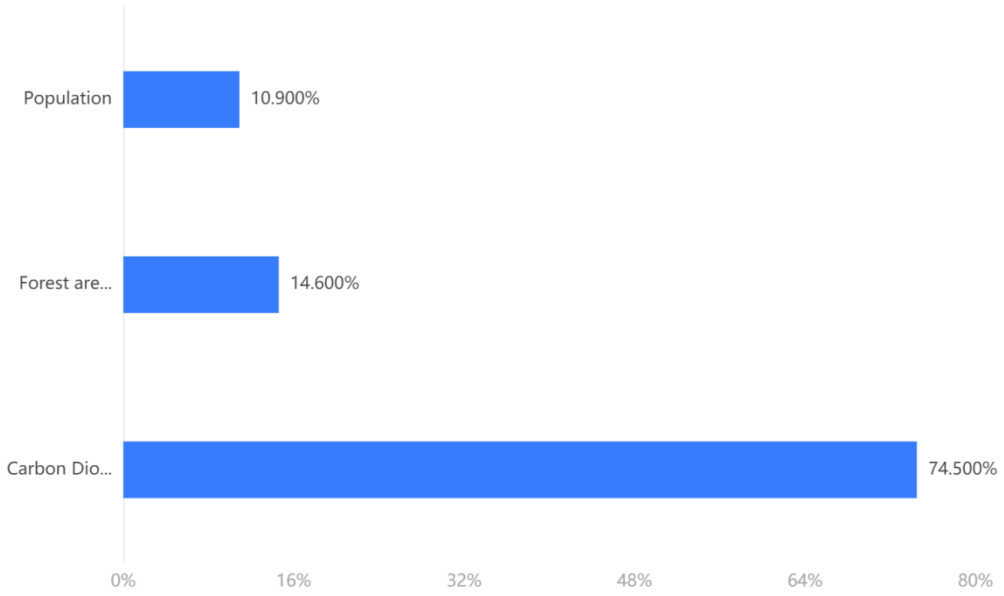

Figure 16 Feature Importance

From Figure 16, we can see that carbon dioxide has the greatest influence on global temperature, and forest area has the second greatest influence on global temperature. Therefore, the main cause of global warming is that the level of carbon dioxide is increasing, while the forest cover is decreasing, which leads to a decrease in the ability to regulate temperature and finally causes global warming.

5.2.5 Curbing or Slowing down Global Warming

The selection of the importance of the characteristics shows that carbon dioxide has the greatest impact on global temperature, while forest area is a close second, so to curb or mitigate global warming mainly from these two aspects.

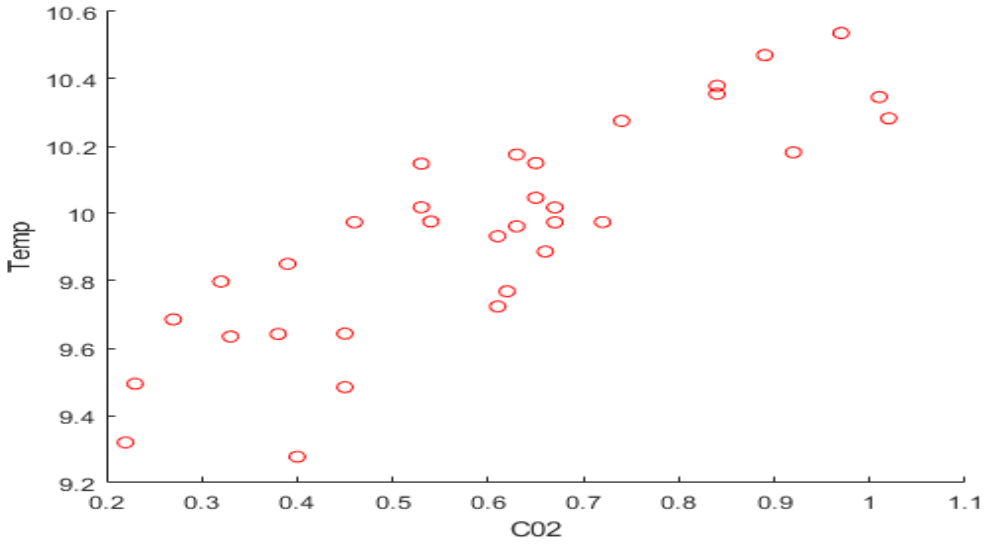

Figure 17 C02 vs. temperature scatter plot

From Figure 17, we can see that when the emission of carbon dioxide is increasing, the global temperature is also increasing. Therefore, we need to reduce the emission of carbon dioxide, try to use some green and clean energy, and reduce the use of coal as well as oil. Improve the level of energy utilization technology and increase the conversion rate of energy so that it can be fully converted, thus reducing the emission of carbon dioxide.

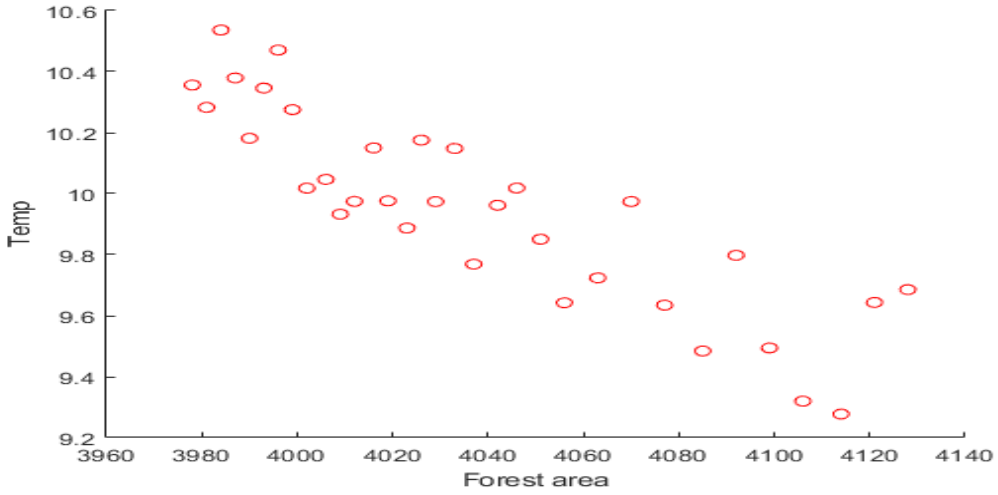

Figure 18 Scatter plot of forest area versus temperature

Figure 18 shows that the global temperature is increasing as the forest area is decreasing. Therefore, if we want to curb or slow down the global warming, we can increase the forest cover by planting trees. We can protect the forest vegetation by not cutting down trees randomly.

5.3 Modeling and Solving Problem 3

The global temperature in March 2022 rises more than during the past decade, with each global temperature rising in a one-point cycle and becoming larger from 1982. Global temperatures are predicted to reach 20 degrees Celsius in 2059 and 22.65 degrees Celsius in 2100.

The global temperature has a slight relationship with geography, with the southern hemisphere having a greater impact on global temperature than the northern hemisphere. Global temperatures are increasing over time. The occurrence of volcanic eruptions and forest fires can cause changes in the structure of the earth's surface, which can affect the climate and ultimately lead to a rise in global temperatures. The main causes of global temperature change are excessive carbon dioxide emissions and decreasing forest cover.

To curb or slow down global warming, the opposite of CO2 emissions should be done by using more green and clean energy and reducing the use of coal. Plant trees, increase forest cover, do not cut down indiscriminately, returning farmland to forest, etc.

6. Evaluation of the Model

6.1 Advantages of the Model

(1) Multi-model combination.
(2) Reliable data.
(3) Modeling and result analysis steps are available for each problem.
(4) The mathematical process is relatively simple and easy to operate.

6.2 Shortcomings of the Model

There is no consideration of the impact of changes in environmental factors.

References

[1] Möbius W, Laan L. Physical and mathematical modeling in experimental papers[J]. Cell, 2015, 163(7): 1577-1583.

[2] Kurasov D. Mathematical modeling system MatLab[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1691(1): 012123.

[3] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science,

2015, 349(6245): 255-260.

[4] Kursa M B, Rudnicki W R. The all relevant feature selection using random forest[J]. arXiv preprint arXiv:1106.5112, 2011.

[5] Qin S J, Chiang L H. Advances and opportunities in machine learning for process data analytics[J]. Computers & Chemical Engineering, 2019, 126: 465-473.